BLIAM: Literature-based Data Synthesis for Synergistic Drug Combination Prediction

Anonymous ACL submission

Abstract

Language models pre-trained on scientific literature corpora have substantially advanced scientific discovery by offering high-quality feature representations for downstream applications. However, these features are often not interpretable, thus can reveal limited insights to domain experts. Instead of obtaining features from language models, we propose BLIAM, a literature-based data synthesis approach to directly generate training data points that are interpretable and model-agnostic to downstream applications. The key idea of BLIAM is to create prompts using existing training data and then use these prompts to synthesize new data points. BLIAM performs these two steps iteratively as new data points will define more informative prompts and new prompts will in turn synthesize more accurate data points. Notably, literature-based data augmentation might introduce data leakage since labels of test data points in downstream applications might have already been mentioned in the language model corpus. To prevent such leakage, we introduce GDSC-combo, a large-scale drug combination discovery dataset that was published after the biomedical language model was trained. We found that BLIAM substantially outperforms a non-augmented approach and manual prompting in this rigorous data split setting. BLIAM can be further used to synthesize data points for novel drugs and cell lines that were not even measured in biomedical experiments. In addition to the promising prediction performance, the data points synthesized by BLIAM are interpretable and model-agnostic, enabling *in silico* augmentations for *in vitro* experiments.

1 Introduction

Scientific literature is an important resource to provide prior knowledge for scientific discovery (Lu Wang et al., 2020; Wei et al., 2013). Natural language processing techniques have been extensively developed to mine scientific literature for scientific problems, including material property prediction (Tshitoyan et al., 2019a), COVID therapeutics research (Lever and Altman, 2021), cancer research (Lever et al., 2019), and drug repurposing (Hsiao and Lu, 2019; Detroja et al., 2022). A promising line of research is to train domain-specific language models on biomedical corpora, which have been used to produce dense representations for downstream supervised learning tasks (Gu et al., 2021; Luo et al., 2022; Lee et al., 2019).

However, these dense representations are not interpretable to domain experts, which is critical for biomedical applications (Stiglic et al., 2020). Although existing interpretable methods can highlight important features within the embeddings (Lundberg and Lee, 2017; Shrikumar et al., 2017), these features are still latent dimensions. Here, we propose to directly synthesize new labeled biological datapoints using pretrained LMs (PLMs). The synthesized data points will be used as augmentations to help train the downstream classifier. A critical issue of synthesizing new data points is the potential data leakage, where test data points might be explicitly mentioned in the scientific literature and thus be seen by PLMs. To address this issue, we introduce a novel large-scale drug combination discovery dataset GDSC-combo (Jaaks et al., 2022) that was published in April 2022 after the corpus that the biomedical PLM used for training (Gu et al., 2021). By examining this dataset in 18 million PubMed abstracts that are seen by the PLM, we found that 99.99% of data points in GDSC-combo never appeared in any abstract. Here, the task is to predict whether a triplet of two drugs and a cell line will be synergistic or not.

We propose the Biomedical Literature Iterative Augmenting Module (BLIAM) to synthesize new triplets from the PLM, which we in turn use to augment the GDSC-combo dataset. BLIAM first mentions sentences that mention two drugs and a cell line in our vocabulary list based on the original dataset (Jaaks et al., 2022). It then clusters these

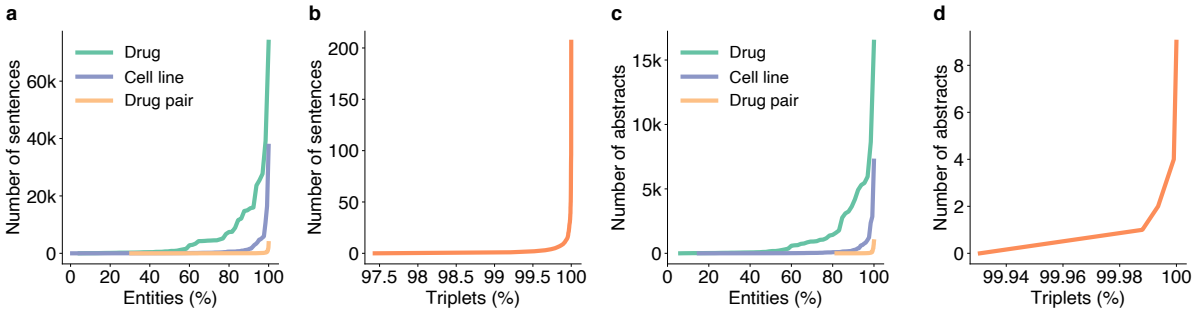

Figure 1: **Examining data leakage by counting the number of sentences and abstracts that mention drugs and cell lines.** Sentences and abstract are obtained from 34,592,344 PubMed papers. x-axis shows the percentage of entities (**a, c**) or triplets (**b, d**) appear in less than k sentences (**a, b**) or abstracts (**c, d**), where k is shown in y-axis. Drug pair stands for two drugs. Triplet stands for two drugs and a cell line.

sentences and derives representative prompts according to cluster centers. These prompts are then used to synthesize new triplets based on the PLM. BLIAM iteratively performs these three steps to create a large and high-quality synthesized synergistic dataset, which is later combined with the original dataset (Jaaks et al., 2022) to train a downstream classifier.

We found that augmenting the GDSC-combo dataset substantially improved downstream synergy prediction from 0.30 AUPRC to 0.38 AUPRC compared to a non-augmentation approach, and that the amount of improvement grows with repeated iterations of BLIAM. BLIAM was particularly beneficial when applied to never-before-seen drugs and cell lines at test time, representing a more realistic setting for real-world clinical applications. In summary, we have introduced a new dataset and task for literature-based data augmentation that does not have the data leakage issue. We have also developed a literature-based data synthesis approach that successfully generates high-quality training data points.

2 Task and Dataset

We focus on drug synergy classification, an important biomedical task to predict whether the combination of two drugs will have a greater-than-additive drug response effect on a given cell line (Jaaks et al., 2022). Predictive models can help fill a large void for synergy screens, as it is too expensive and time-consuming to exhaustively search the combinations of drugs and cell lines *in vitro*. However, the sparsity of available labeled drug synergy datasets also makes this a challenging problem, and effectively leveraging prior knowledge through existing literature represents one possible route toward improved synergy prediction.

We formulate the problem of drug synergy prediction as a binary classification task. We obtained an initial set of input triplets and corresponding synergy labels $\mathcal{D}_{full} = \{((d_i^A, d_i^B, c_i), y_i)\}_{i=1}^{N'}$ from a recent dataset (Jaaks et al., 2022), where $y_i \in \{0, 1\}$ indicates whether drugs d_i^A and d_i^B are synergistic in cell line c_i. Based on this initial set, we could then learn the classifier $f_\theta(d_i^A, d_i^B, c_i) \in [0, 1]$, which reflects the probability that a triplet is synergistic.

We further filtered the dataset $\mathcal{D}_{full}$ using the PubMedBERT (Gu et al., 2021) vocabulary $\mathcal{V}$, where we only retained triplets with $d_i^A \in \mathcal{V} \vee d_i^B \in \mathcal{V} \vee c_i \in \mathcal{V}$, producing a reduced dataset $\mathcal{D}$ of size $N = 24{,}890$, where only 191 triplets have synergistic effect $y_i = 1$. In total, $\mathcal{D}$ contains 125 unique cell lines and 65 unique drugs. In this work, we do not consider additional classification features for drugs or cell lines, such as gene expression or drug SMILES representation. However, our framework can also incorporate such features as long as we can obtain these features from other databases after we have synthesized the triplets.

Our goal is to use the pretrained PubMedBERT (Gu et al., 2021) to generate a synthetic dataset $\tilde{\mathcal{D}}$ and learn the classifier $f_\theta(d_i^A, d_i^B, c_i) \in [0, 1]$ using $\tilde{\mathcal{D}} \cup \mathcal{D}$. Importantly, the drug synergy dataset from Jaaks et al. was published after the PLM from Gu et al., enabling us to avoid data leakage between the PLM and the synergy classification data splits. **Fig. 1** shows the proportion of entities and their appearance in sentences and abstracts. Despite single drugs, single cell lines, and drug pairs frequently appearing in sentences and abstracts, triplets exhibit low occurrence in PubMed literature, further confirming that using PubMedBERT to augment

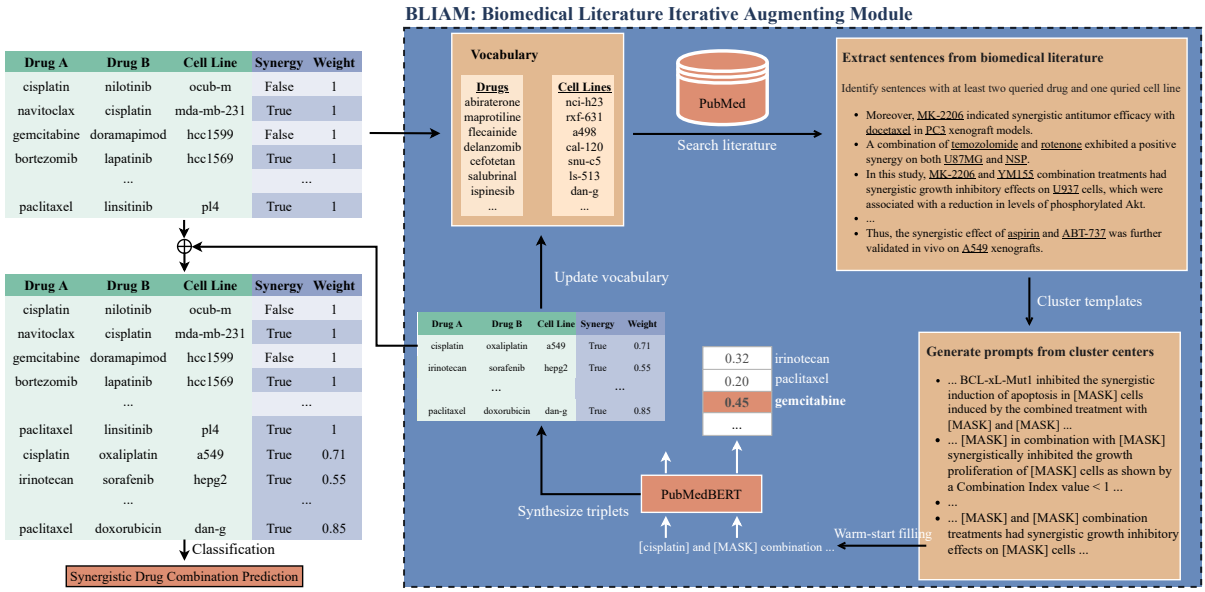

Figure 2: **Overview of BLIAM.** BLIAM begins by searching PubMed literature using an initial vocabulary of drugs and cell lines. The sentences obtained are then clustered and medoids of each cluster are used to create prompts for PubMedBERT for cloze prompt filling. Tokens filled in the drug and cell line masks are used to expand the initial vocabulary, even if they are not valid drug and cell line names. These steps (blue box) are performed iteratively. After 3 iterations, generated triplets are used as an augmentation dataset for synergy prediction task.

GDSC-combo does not bring the risk of data leakage.

3 Methods

3.1 Base classifier

For a given triplet $(d_i^A, d_i^B, c_i) \in \mathcal{D}$, we aim to learn a classifier $f_\theta(d_i^A, d_i^B, c_i) \rightarrow [0, 1]$. In particular, we design a classifier consisting of two embedding matrices for drugs and cell lines. We didn't use additional features for drugs or cell lines so the embedding matrices are randomly initialized. The embedding vectors for three entities from the triplet are concatenated and passed into a multi-layer neural network with Leaky ReLU activation functions for classification. We followed previous work on drug synergy prediction for the neural network architecture (Preuer et al., 2018).

When augmented data is available, we combine the drug synergy dataset $\mathcal{D} = \{((d_i^A, d_i^B, c_i), y_i)\}_{i=1}^N$ with the augmented synergistic dataset $\tilde{\mathcal{D}} = \{((\tilde{d_j^A}, \tilde{d_j^B}, \tilde{c_j}), 1)\}_{j=1}^M$ and learn a new classifier $f'_\theta(d^A, d^B, c) \in [0, 1]$. We train the model with binary cross entropy as $\mathcal{L}(\theta|\mathcal{D}, \tilde{\mathcal{D}}) = \sum_{i=1}^N \texttt{BCE}(y_i, f'_\theta(d_i^A, d_i^B, c_i)) + \sum_{j=1}^M w_j \log(f'_\theta(\tilde{d_j^A}, \tilde{d_j^B}, \tilde{c_j}))$, where BCE indicates binary cross entropy loss and w_j is a weight factor according to the likelihood of each synthetic triplet generated by PLM. At inference time, we predict a triplet to be synergistic if $f'_\theta(d^A, d^B, c) > 0.5$.

3.2 Manual Prompts

We first develop a few manual prompts to augment the data. We manually construct a list of cloze prompts that can be filled by PubMedBERT (see **Supplementary A.1**), such as `On cell line [MASK], [MASK] has synergy with [MASK] and [MASK] and [MASK] are effective to treat cell line [MASK]`. We then take the most likely token output by the PLM for each masked token to construct new drug combinations and cell line triplets, to which we assign positive $y = 1$ synergy labels. We denote the union of the original dataset and this simulated dataset as $\tilde{\mathcal{D}}_{man}$.

3.3 Literature-mined prompts

Although manual prompts are more intuitive, designing good templates requires domain knowledge and may not cover all aspects of a specific task. To address this, we extract sentences mentioning drugs and cell lines from scientific literature, which serve as candidates for automated prompt generation (**Fig. 2**).

To search for candidate templates, we first build a larger vocabulary of single drugs and single cell lines by combining the set of all drugs and cell lines

from our dataset $\mathcal{D}$ with the drugs and cell lines from some existing databases, including LINCS (Subramanian et al., 2017), Genomics of Drug Sensitivity in Cancer (GDSC) (Iorio et al., 2016), Cancer Cell Line Encyclopedia (CCLE) (Meyers et al., 2017; Tsherniak et al., 2017), and National Cancer Institute 60 human cell line (NCI-60) (Shoemaker, 2006). This increases the number of drugs from 125 to 1,486 and the number of cell lines from 65 to 2,037, ensuring that our collection of drugs and cell lines is large enough to identify diverse templates. This step does not introduce data leakage as these external databases are not drug combination databases and we only collect drug and cell line names from them.

We then search within the PubMed abstract collection (Wei et al., 2019) for sentences containing at least two drugs and one cell line, which can therefore be interpreted as a triplet. After obtaining this set of sentences, we then filter them by keywords such as *synergy* and *synergistic* to ensure that the sentence relates to synergy.

We can then mask the triplets within each sentence to obtain a prompt. However, as treating every resulting sentence as a prompt would be time-consuming and redundant, we only retain representative prompts. Specifically, we masked all drug and cell line mentioned within a sentence and computed its BERT embedding. Then, we compute the k-medoid clustering of the masked sentence embeddings. Finally, we selected the medoid of each cluster to use as the cluster's representative prompt.

3.4 Prompt warm-start filling

We found that directly filling in three masked tokens in a prompt could generate many redundant triplets, even using different prompts, thus leading to less efficient generation. To address this issue, we have developed a warm-start filling strategy. Specifically, we uniformly sample a triplet from the current dataset and uses one or two elements in this triplet to pre-fill the prompt. This strategy will avoid frequently filling the prompt with popular drugs or cell lines, thus enabling us to synthesize triplets containing rare cell line or drugs and improving downstream classification results.

Since some triplets might contain drug or cell line name that is not in the PubMedBERT vocabulary, we further restrict to only warm-start fill in-vocabulary entity. For instance, given the triplet (`Cisplatin`, `Camptothecin`, `BT-483`) and prompt template `[MASK] and [MASK] combination treatments had synergistic growth inhibitory effects on [MASK] cells`, Cisplatin is the only triplet element that appears in the PubMedBERT vocabulary. Therefore, we could pre-fill the prompt template as `Cisplatin and [MASK] combination treatments had synergistic growth inhibitory effects on [MASK] cells`, or `[MASK] and cisplatin combination treatments had synergistic growth inhibitory effects on [MASK] cells`. All token filling are type-aware: a drug (cell line) mask will only be replaced with a drug (cell line) name. In the case where the prompt template contains additional `[MASK]` tokens (i.e., more than two drug mask positions or more than one cell line mask position), all masked locations are considered possible options for warm-start filling.

We then use these warm-start filled prompts, all containing at least one remaining masked token, to further prompt the PLM. In the case of more than three `[MASK]` tokens, we enumerate all combinations of two drugs and one cell line to form new triplets. This warm-start filling strategy helps us achieve better diversity while still consider contextual information.

3.5 Iterative augmentation

Furthermore, we repeat the literature-mined prompting process multiple times to iteratively construct the augmented dataset (**Fig. 2**). Iteratively adding new triplets to the dataset enables BLIAM to expand its set of possible prompts based on previous literature-based findings. This can both increase the potential diversity of prompts and improve prompting quality, with subsequent iterations able to learn from the synthesized data triplets from previous iterations. Given the initial dataset $\mathcal{D}$, we fill literature-mined prompts to produce a new triplet dataset $\tilde{\mathcal{D}}_{liter}$. We then mine literature again, including triplets from $\mathcal{D} \cup \tilde{\mathcal{D}}_{liter}$, fill the resulting prompts, and query PubMedBERT to produce the newly augmented triplet dataset $\tilde{\mathcal{D}}^{(2)}_{liter}$. We repeat this process for a total of γ iterations, using the simulated triplets from the previous iteration to help inform the prompts for the next, and denote the resulting iterated dataset as $\tilde{\mathcal{D}}_{it} = \mathcal{D} \cup \tilde{\mathcal{D}}^{(3)}_{liter}$.

3.6 Vocabulary-restricted augmentation

Although we provide context to PubMedBERT in the form of the filled-in prompts, it is possible that

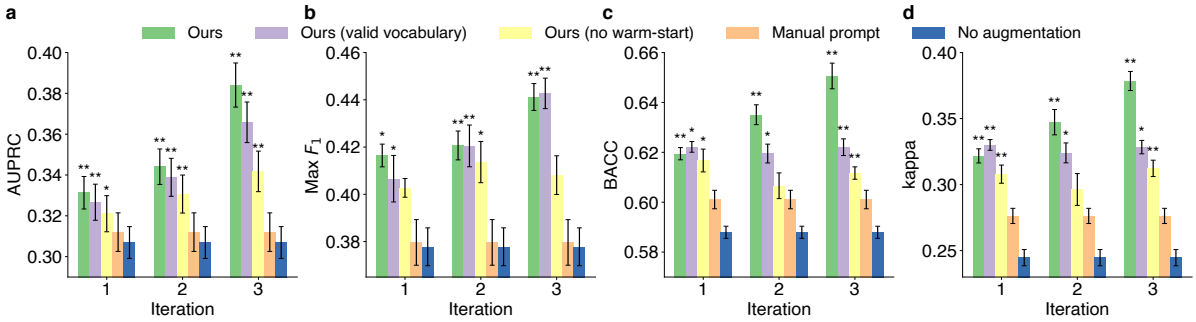

Figure 3: **Performance on downstream drug synergistic prediction evaluated using four metrics.** Vertical bars represent standard error bars. Stars above bars indicate statistical significance when comparing with *No augmentation* setting using one-sided paired t-tests ($*$: $p < 0.05$, $**$: $p < 0.01$). y-axis shows the iteration. Ours, Ours (valid vocabulary), Manual prompt, No augmentation stand for training on $\tilde{\mathcal{D}}_{it}, \tilde{\mathcal{D}}_{valid}, \tilde{\mathcal{D}}_{manul}$, and $\mathcal{D}$, respectively. Ours (no warm-start) stands for our methods without warm-start filling.

the most likely generated words for the [MASK] tokens do not correspond to actual drugs or cell lines. We address this in one variant of our method, where we restrict the set of generated tokens we consider to be within the set of drugs or cell lines from $\mathcal{D}$, LINCS, GDSC, CCLE, and NCI-60 for drug and cell line mask masks, respectively. This ensures that the generated triplets do in fact correspond to two drugs and a cell line. This additional filtering step makes the synthesized triplets and consequent classifier performance more interpretable and also enables biologists to inspect and understand BLIAM's synthesized dataset. We then repeat the iterative process from Section 3.5 and refer to the resulting iterated dataset, now restricted to valid triplets, as $\tilde{\mathcal{D}}_{valid}$.

4 Experimental setting

We chose PubMedBERT-abstract[1] as our PLM. After intersecting the dataset with PubMedBERT's vocabulary, there are 24,890 instances, 191 of which have a positive label. We repeated the recursive search in literature-mined prompts for $\gamma = 3$ iterations.

We conducted the triplet synergy classification task through 5-fold stratified cross-validation. We trained the classifier using the Adam optimizer with $\epsilon = 10^{-8}, \beta_1 = 0.9, \beta_2 = 0.999$ for 50 epochs. We set the batch size to 64. We performed grid search on learning rates over $[0.01, 0.005, 0.001]$, hidden dimensions over $[128, 256, 512]$, and warm-up epochs over $[0, 5, 10, 20, 30, 40]$. The maximum number of parameters of the classifier is 343,874.

[1] https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract, Last accessed: December 11, 2022.

For training with the augmentation dataset, we considered whether using instance weights on loss from synthetic instances as an additional hyperparameter. All experiments were conducted on Nvidia RTX 3090 GPU with 24GB Memory. Training and inference can be finished within one GPU hour. We followed previous works to adopt four metrics that can evaluate the imbalanced classification task (Lin et al., 2022): AUPRC, max F1 score, balanced accuracy (BACC), and Cohen's kappa (kappa).

5 Experimental results

5.1 BLIAM-derived augmentations improve downstream classification

We first investigated whether the BLIAM-derived augmentations can improve downstream drug synergy classification. We compared the performance of training using the unaugmented dataset $\mathcal{D}$ with the three variants of PLM-augmented datasets: $\tilde{\mathcal{D}}_{it}$ and $\tilde{\mathcal{D}}_{valid}$ produced by BLIAM, and $\tilde{\mathcal{D}}_{man}$ from the manual prompting. We found that all three augmented datasets outperformed the unaugmented dataset $\mathcal{D}$ (**Fig. 3**), indicating the effectiveness of using literature to generate high-quality biomedical training data points.

Among these three augmented datasets, the $\tilde{\mathcal{D}}_{it}$ augmented dataset led to the best overall performance, with an AUPRC of 0.38, max F_1 score of 0.44, BACC of 0.65, and kappa of 0.38, relative to the 0.31, 0.38, 0.59, and 0.25 respectively for the unaugmented $\mathcal{D}$. $\tilde{\mathcal{D}}_{valid}$ restricts the prompts to only synthesize valid triplets of drugs and cell lines, thus providing interpretable triplets for domain experts. We found that the performance of

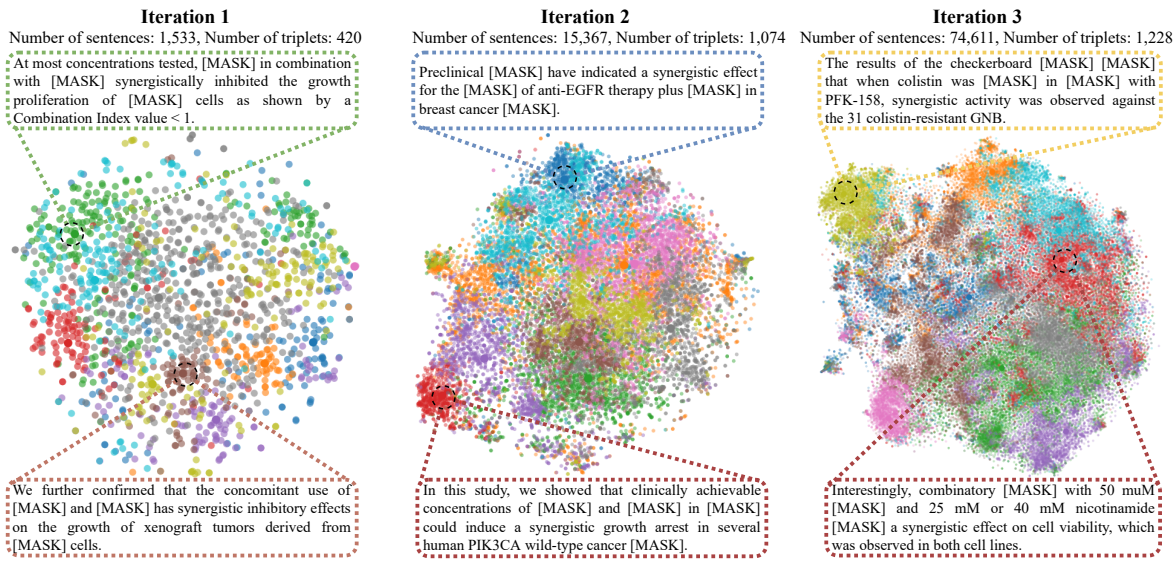

Figure 4: **t-SNE plots on literature-mined sentences in three iterations.** Sentences are clustered into ten clusters in each iteration. Each sentence is colored according to its belonging cluster. One prompt template is derived according to the center of each cluster.

training on $\tilde{\mathcal{D}}_{valid}$ is slightly worse than $\tilde{\mathcal{D}}_{it}$, suggesting a trade-off between restricted interpretable results and downstream prediction performance. Nevertheless, the performance of $\tilde{\mathcal{D}}_{valid}$ is still better than $\tilde{\mathcal{D}}_{man}$ and $\mathcal{D}$, again confirming the benefit of our iterative augmentation approach. We also observed that our method outperformed the variant with warm-start, demonstrating the effectiveness of warm-start in prompt filling.

Moreover, we noticed that the performance of $\tilde{\mathcal{D}}_{it}$ improves over iterations. For example, the AUPRC of $\tilde{\mathcal{D}}_{it}$ improves from 0.33 to 0.34 after the second iteration and further boosts to 0.38 after the third iteration. Since our approach generates new prompts according to the data points synthesized at each step, these new prompts are learned using a larger data set and thus might be more robust. We did not observe further improvement after more than three iterations. Collectively, the improvement of our approach against the unaugmented approach and manual prompting demonstrates the preeminence of iteratively using a PLM to synthesize new data points.

5.2 Iterative augmentation expands and improves the augmented triplets

To further investigate the benefit of gradually expanding the augmented dataset with repeated iterations of the augmenting module, we inspected the prompts identified at each of our three iterations. Manual prompts identified a total of 636 simulated triplets. By contrast, the first iteration of literature-based prompting led to 420 simulated triplets, which increased to $1{,}074$ and $1{,}228$ simulated triplets in the second and third iterations, respectively, suggesting larger and more robust augmentations. These triplets were also generated from an increasing number of sentences (**Fig. 4**). We further examined the t-Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) representations of sentences extracted at each iteration (**Fig. 4**). Here, the prompt template is derived from the center of each cluster. As the set of prompt templates expands over iterations, we found that the clustering pattern of the sentences becomes more visible. This in turn suggests that the clustering for prompt template selection becomes more robust at later iterations of the augmenter. Interestingly, we found that the prompt templates at later iterations are less compatible with human intuition. For example, despite the existence of drug-related context, templates in the third iteration do not seem to mask a drug or cell line token. This observation is consistent with previous works (Shin et al., 2020) where automatically derived prompts often contain non-informative tokens. Nevertheless, although the prompt templates might not be intuitive to human experts, the generated triplets are interpretable and can help biologists identify new synergistic triplets.

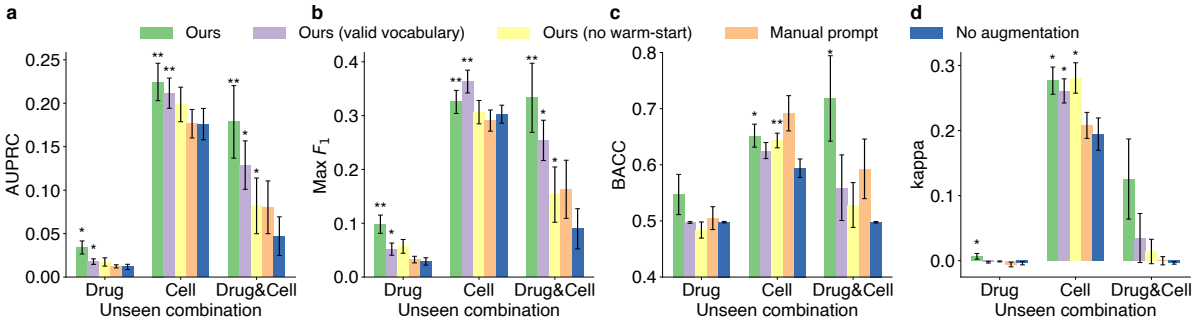

Figure 5: **Performance on triplets that contain never-before-seen drugs or cell lines evaluated using four metrics.** Vertical bars represent standard error bars. Stars above bars indicate statistical significance when comparing with *No augmentation* setting using one-sided paired t-tests ($*$: $p < 0.05$, $**$: $p < 0.01$). Ours, Ours (valid vocabulary), Manual prompt, No augmentation stand for training on, $\tilde{\mathcal{D}}_{it}$, $\tilde{\mathcal{D}}_{valid}$, $\tilde{\mathcal{D}}_{manul}$, and $\mathcal{D}$, respectively. Ours (no warm-start) stands for our methods without warm-start filling. Performance of iteration 3 is shown here.

5.3 BLIAM synthesizes triplets that contain never-before-seen drugs and cell lines

A critical limitation of experimentally-derived drug synergistic datasets is the existence of new drugs and cell lines at inference time, which hinders the progress of predicting patient outcomes in real-world clinics. Incorporating prior knowledge from biomedical literature might help to address this problem. To investigate our ability to leverage the PLM to classify triplets that contain never-before-seen drugs and cell lines, we designed three new test settings: *Drug*, *Cell*, and *Drug&Cell*. Given a set of test triplets $\{(d_i^A, d_i^B, c_i)\}_{i=1}^T$, we define each setting as follows: in *Drug*, we exclude any triplets containing d_i^A or d_i^B from the input training dataset $\mathcal{D}$ for all i; in *Cell*, we exclude any triplets containing c_i from $\mathcal{D}$ for all i; in *Drug&Cell*, we exclude any triplets containing d_i^A, d_i^B, or c_i from $\mathcal{D}$ for all i.

In these challenging but more realistic settings, we found that BLIAM was again beneficial for synergistic classification, relative to no augmentation or manual prompting (**Fig. 5**). For example, in *Cell*, the AUPRC was 0.18 without augmentation or with manual prompting, and increased to 0.22 with BLIAM's augmented dataset. The AUPRC was also better if we restricted the vocabulary (0.21) or even skipped the warm-start filling (0.20). The improvement was even larger on the most challenging setting of *Drug&Cell*, where the unaugmented $\mathcal{D}$ had an AUPRC of 0.05, $\tilde{\mathcal{D}}_{man}$ had an AUPRC of 0.08, and our $\tilde{\mathcal{D}}_{it}$ had an AUPRC of 0.18. While *in vitro* experimental approaches might be too expensive or otherwise infeasible to measure certain drugs and cell lines, BLIAM complements them by synthesizing high-quality training data points, paving the way for applying these experimental measurements to real-world clinics.

6 Related Work

6.1 PLMs for data augmentation

Recent works have shown that large PLMs can function as good few-shot and zero-shot learners (Brown et al., 2020; Radford et al., 2019), and thoughtful probing of PLMs has been used to improve downstream model performance in low-data settings (Zhong et al., 2021). PLMs have further been used for data augmentation, including word-substitution-based approaches (Hu et al., 2019), sentence-substitution (Kumar et al., 2019), back-translation (Xie et al., 2020), prompting (Wang et al., 2022; Shin et al., 2020; Schick and Schütze, 2021), and textual generation for synthetic dataset construction (Yang et al., 2020; Mekala et al., 2022; Zhou et al., 2022; Puri et al., 2020; Vu et al., 2021; Meng et al., 2022; Ye et al., 2022; Yoo et al., 2021; Lewis et al., 2019; Wei et al., 2021). In particular, PLM-based augmentation techniques have improved downstream model performance in imbalanced classification settings (Liu et al., 2020), where generating synthetic samples for the minority class improves performance (Liu et al., 2020; Chawla et al., 2002).

Most similar to this work, several data augmentation techniques use cloze test formulations (Chapelle and Abraham, 1990), where the PLM is used in a fill-in-the-blank task (Ng et al., 2020; Cai et al., 2020; Schick and Schütze, 2021) to augment a textual corpus. In this work, however, we use a cloze test formulation to synthesize novel, in-

terpretable data points for downstream biomedical applications. The generated tokens are therefore taken out of their surrounding textual context after the iterative dataset synthesis is complete. Moreover, we prevent data leakage by carefully designing our task.

6.2 PLMs as relational extractors

PLMs have also been shown to function as good relational extractors, effectively distilling information contained in the pretraining textual corpus even in the few-shot setting (Brown et al., 2020). Several works have found that using cloze sentence formulations with PLMs can behave as relation extractors (Petroni et al., 2019; Jiang et al., 2020), presenting an alternative to large knowledge bases (Baldini Soares et al., 2019; Shwartz et al., 2020). Most similar to our work, Bouraoui et al. use a known relational pair to identify relevant templates and use a PLM to predict other tokens that may belong to the same template relation. In this work, however, we study how literature and PLM can improve downstream application that does not involve text data.

6.3 Literature mining for scientific discovery

NLP and literature-mining techniques have proven to be a useful technique to digest vast amounts of scientific and medical text and extract useful knowledge. Downstream applications include material science property prediction (Tshitoyan et al., 2019b), clinical outcome forecasting (Naik et al., 2022; van Aken et al., 2021; Jin et al., 2021), yeast cell-signaling modeling (Coutant et al., 2019), cancer gene identification (Hsiao and Lu, 2019; Lever et al., 2019), and drug repurposing (Nye et al., 2021; Jin et al., 2021; Tworowski et al., 2021). One work uses PLMs for a biomedical knowledge graph (KG) completion task (Nadkarni et al., 2021), though does not focus on downstream applications of such KGs as what we did. Recently, Shim et al. employed document-based feature extraction to improve anti-cancer drug synergy prediction performance (Shim et al., 2022). In this work, we build upon these ideas and use a PLM to synthesize interpretable triplets for dataset augmentation.

7 Conclusion and Future work

In this work, we have demonstrated that iteratively prompting PLMs with literature-mined templates produces interpretable data points that improve performance on a downstream drug synergy prediction task. We found that BLIAM's augmented dataset outperforms no augmentation and manual prompting, even though the prompts themselves are not always readily interpretable. We further found that BLIAM-synthesized triplets were particularly useful for drugs and cell lines not present in the original biomedical dataset, but which might have some literature evidence distilled by the PLM.

Although literature-mined prompts for PLMs improve our drug synergy classification performance, there are some direct extensions to the methodology that could further improve and refine the augmented dataset. First, when searching for sentences that contain two drugs and a cell line, we do not consider the overall intent of the sentence. For instance, we do not investigate negation before assigning a positive synergy label (e.g., `[MASK] and [MASK] are` *`not`* `synergistic in cell line [MASK]`), or general research statements that are not intended to convey a result (e.g., `In this work, we investigated the possible synergy of [MASK] and [MASK] in [MASK] models`). Future work could therefore combine a sentiment analysis approach with our prompt construction to identify templates that truly imply a discovered synergistic triplet. Second, we limit the diversity of our generated triplets by always selecting the most likely generated word at each of our masked tokens. Instead, future work could expand the number of generated triplets by sampling from the generated distribution and continuing to weight the augmented dataset according to triplet likelihood in the classification loss to minimize the impact of false positives.

Furthermore, LMs pretrained on biomedical corpora could be applied more directly to biomedical tasks. Rather than constructing an augmented triplet dataset for drug synergy prediction, a PLM could be directly queried with `Cisplatin and [MASK] are synergistic in cell line dan-g` to produce a distribution over possible secondary drugs and directly answer the drug synergy question. Moreover, such approaches can be combined with recent prompting techniques such as chain-of-thought prompting (Wei et al., 2022). Building upon recent work (Liang et al., 2022), future studies could also combine other modalities to convey additional relevant biomedical information to the PLM, such as the molecular structure of a drug or marker genes for the cell line.

8 Limitations

Although we found that BLIAM improved downstream drug synergy prediction, there are several important limitations. First, the prompt template quality is dependent on the literature sentences containing the biomedical entities of interest. Second, the iterative prompting procedure could amplify inaccurate triplets. Third, the conditions measured in biomedical datasets and literature may contain their own biases, which are in turn internalized by the PLM and amplified by the iterative prompter. Finally, the diversity of generated triplets is restricted to PLM used. Although we can restrict the output to valid drugs and cell lines, it is unknown whether the drug pairs from the synthetic triplet are indeed synergistic on the cell line.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. *Proc. Conf. AAAI Artif. Intell.*, 34(05):7456–7463.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, number Article 159 in NIPS'20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343, Online. Association for Computational Linguistics.

Carol A Chapelle and Roberta G Abraham. 1990. Cloze method: what difference does it make? *Language Testing*, 7(2):121–146.

N V Chawla, K W Bowyer, L O Hall, and W P Kegelmeyer. 2002. SMOTE: Synthetic minority oversampling technique. *J. Artif. Intell. Res.*, 16:321–357.

Anthony Coutant, Katherine Roper, Daniel Trejo-Banos, Dominique Bouthinon, Martin Carpenter, Jacek Grzebyta, Guillaume Santini, Henry Soldano, Mohamed Elati, Jan Ramon, Celine Rouveirol, Larisa N Soldatova, and Ross D King. 2019. Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 116(36):18142–18147.

Trishna Saha Detroja, Hava Gil-Henn, and Abraham O Samson. 2022. Text-Mining approach to identify hub genes of cancer metastasis and potential drug repurposing to target them. *J. Clin. Med. Res.*, 11(8).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1):1–23.

Yi-Wen Hsiao and Tzu-Pin Lu. 2019. Text-mining in cancer research may help identify effective treatments. *Transl Lung Cancer Res*, 8(Suppl 4):S460–S463.

Zhiting Hu, Bowen Tan, Ruslan Salakhutdinov, Tom Mitchell, and Eric P Xing. 2019. Learning data manipulation for augmentation and weighting. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA.

Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K Egan, Qingsong Liu, Tatiana Mironenko, Xeni Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S Gray, Daniel A Haber, Michael R Stratton, Cyril H Benes, Lodewyk F A Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J Garnett. 2016. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754.

Patricia Jaaks, Elizabeth A Coker, Daniel J Vis, Olivia Edwards, Emma F Carpenter, Simonetta M Leto, Lisa Dwane, Francesco Sassi, Howard Lightfoot, Syd Barthorpe, Dieudonne van der Meer, Wanjuan Yang, Alexandra Beck, Tatiana Mironenko, Caitlin Hall, James Hall, Iman Mali, Laura Richardson, Charlotte Tolley, James Morris, Frances Thomas, Ermira Lleshi, Nanne Aben, Cyril H Benes, Andrea Bertotti, Livio Trusolino, Lodewyk Wessels, and Mathew J Garnett. 2022. Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature*, 603(7899):166–173.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Trans. Assoc. Comput. Linguist.*, 8:423–438.

Shuting Jin, Zhangming Niu, Changzhi Jiang, Wei Huang, Feng Xia, Xurui Jin, Xiangrong Liu, and Xiangxiang Zeng. 2021. HeTDR: Drug repositioning based on heterogeneous networks and text mining. *Patterns (N Y)*, 2(8):100307.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jake Lever and Russ B. Altman. 2021. Analyzing the vast coronavirus literature with coronacentral. *Proceedings of the National Academy of Sciences of the United States of America*, 118.

Jake Lever, Eric Y Zhao, Jasleen Grewal, Martin R Jones, and Steven J M Jones. 2019. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods*, 16(6):505–507.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910, Florence, Italy. Association for Computational Linguistics.

Sheng Liang, Mengjie Zhao, and Hinrich Schuetze. 2022. Modular and Parameter-Efficient multimodal fusion with prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2976–2985, Dublin, Ireland. Association for Computational Linguistics.

Jiacheng Lin, Hanwen Xu, Addie Woicik, Jianzhu Ma, and Sheng Wang. 2022. Pisces: A cross-modal contrastive learning approach to synergistic drug combination prediction.

Ruibo Liu, Guangxuan Xu, and Soroush Vosoughi. 2020. Enhanced offensive language detection through data augmentation. *ArXiv*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The covid-19 open research dataset. *ArXiv*.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.*, 23(6):bbac409.

L V D Maaten and Geoffrey E Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*.

Dheeraj Mekala, Tu Vu, Timo Schick, and Jingbo Shang. 2022. Leveraging QA datasets to improve generative data augmentation.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*.

Robin M Meyers, Jordan G Bryan, James M McFarland, Barbara A Weir, Ann E Sizemore, Han Xu, Neekesh V Dharia, Phillip G Montgomery, Glenn S Cowley, Sasha Pantel, Amy Goodale, Yenarae Lee, Levi D Ali, Guozhi Jiang, Rakela Lubonja, William F Harrington, Matthew Strickland, Ting Wu, Derek C Hawes, Victor A Zhivich, Meghan R Wyatt, Zohra Kalani, Jaime J Chang, Michael Okamoto, Kimberly Stegmaier, Todd R Golub, Jesse S Boehm, Francisca Vazquez, David E Root, William C Hahn, and Aviad Tsherniak. 2017. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.*, 49(12):1779–1784.

R Nadkarni, D Wadden, I Beltagy, N A Smith, and others. 2021. Scientific language models for biomedical knowledge base completion: an empirical study. *arXiv preprint arXiv*.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-Augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States. Association for Computational Linguistics.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-Supervised manifold based data augmentation for improving Out-of-Domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Benjamin E Nye, Jay DeYoung, Eric Lehman, Ani Nenkova, Iain J Marshall, and Byron C Wallace. 2021. Understanding clinical trial reports: Extracting medical entities and their relations. *AMIA Jt Summits Transl Sci Proc*, 2021:485–494.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Kristina Preuer, Richard P I Lewis, Sepp Hochreiter, Andreas Bender, Krishna C Bulusu, and Günter Klambauer. 2018. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, 34(9):1538–1546.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, D Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical Report*.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Yongsun Shim, Munhwan Lee, Pil-Jong Kim, and Hong-Gee Kim. 2022. A novel approach to predicting the synergy of anti-cancer drug combinations using document-based feature extraction. *BMC Bioinformatics*, 23(1):163.

Taylor Shin, Yasaman Razeghi, Robert L Logan, IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Robert H Shoemaker. 2006. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, 6(10):813–823.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with Self-Talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 10(5):e1379.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, David L Lahr, Jodi E Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M Enache, Federica Piccioni, Sarah A Johnson, Nicholas J Lyons, Alice H Berger, Alykhan F Shamji, Angela N Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S Gray, Paul A Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J Haggarty, Lucienne V Ronco, Jesse S Boehm, Stuart L Schreiber, John G Doench, Joshua A Bittker, David E Root, Bang Wong, and Todd R Golub. 2017. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, 171(6):1437–1452.e17.

Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, Robin M Meyers, Levi Ali, Amy Goodale, Yenarae Lee, Guozhi Jiang, Jessica Hsiao, William F J Gerath, Sara Howell, Erin Merkel, Mahmoud Ghandi, Levi A Garraway, David E Root, Todd R Golub, Jesse S Boehm, and William C Hahn. 2017. Defining a Cancer Dependency Map. *Cell*, 170(3):564–576.e16.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alex Dunn, Ziqin Rong, Olga Vitalievna Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019a. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571:95–98.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019b. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Dmitry Tworowski, Alessandro Gorohovski, Sumit Mukherjee, Gon Carmi, Eliad Levy, Rajesh Detroja, Sunanda Biswas Mukherjee, and Milana Frenkel-Morgenstern. 2021. COVID19 drug repository: text-mining the literature in search of putative COVID19

therapeutics. *Nucleic Acids Res.*, 49(D1):D1113–D1121.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. Clinical outcome prediction from admission notes using Self-Supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-Training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. PromDA: Prompt-based data augmentation for Low-Resource NLU tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4255, Dublin, Ireland. Association for Computational Linguistics.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, 47(W1):W587–W593.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 41:W518 – W522.

Jason Wei, Chengyu Huang, Soroush Vosoughi, Yu Cheng, and Shiqi Xu. 2021. Few-Shot text classification with triplet networks, data augmentation, and curriculum learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5493–5500, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. *ArXiv*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are Human-Level prompt engineers.

A Appendix

A.1 Manual Prompts

- On cell line [MASK], [MASK] has synergy with [MASK].
- On cell line [MASK], [MASK] are synergistic with [MASK].
- [MASK] has synergy with [MASK] on cell line [MASK].
- [MASK] and [MASK] are synergistic on cell line [MASK].
- On cell line [MASK], there is a synergy between [MASK] and [MASK].
- There is a synergy between [MASK] and [MASK] on cell line [MASK].
- [MASK] and [MASK] are effective to treat to cell line [MASK].
- [MASK] and [MASK] are effective on cell line [MASK].
- On cell line [MASK], [MASK] and [MASK] are effective.
- On cell line [MASK], [MASK] and [MASK] are synergistic.
- On cell line [MASK], [MASK] and [MASK] have an synergy.

A.2 Literature-mined prompts

A.2.1 Iteration 1

- As shown in Figure 4G, overexpression of BCL-xL or, to a slightly lesser extent, BCL-xL-Mut1 inhibited the synergistic induction of apoptosis in [MASK] cells induced by the combined treatment with [MASK] and [MASK].
- We observed synergistic effects on [MASK] cells as well as three additional gastric cancer cell lines with FGFR2 amplification when [MASK] was combined with small molecular inhibitors Cpd22 and [MASK] targeting ILK and EGFR/HER2, respectively.
- At most concentrations tested, [MASK] in combination with [MASK] synergistically inhibited the growth proliferation of [MASK] cells as shown by a Combination Index value < 1.
- Combination treatment with [MASK] and [MASK] caused synergistically increased cell death in [MASK] and [MASK] cells.
- In this study, [MASK] and [MASK] combination treatments had synergistic growth inhibitory effects on [MASK] cells, which were associated with a reduction in levels of phosphorylated Akt.
- We further confirmed that the concomitant use of [MASK] and [MASK] has synergistic inhibitory effects on the growth of xenograft tumors derived from [MASK] cells.
- Aneu-MKN45 developed a resistance to [MASK] which could be reversed by HZ08; Flow cytometry and western-blot indicates that HZ08-combination could induce apoptosis and increase the expression of apoptosis-related biomarkers on aneu-MKN45; in vivo study also reflect the same correlation between aneuploidy and cisplatin-resistance, which could be antagonized by HZ08 combination; When investigating the involved pathway, in anue-MKN45, the expression of molecules in p53 pathway was decreased; HZ08 could increase the expression of p53 down-stream molecules as well as elevate the activity of p53, while inhibiting Mdm2, the major negative regulator of p53; p53 inhibitor [MASK] could completely abrogate HZ08 's synergism effects, and mimic cisplatin-resistance on dip-MKN45.Lower p53 pathway expression that attenuates cisplatin-induced apoptosis might be at least partly the reason of cisplatin-resistance occurred in aneuploid [MASK] both in vitro and in vivo; Combination of HZ08 could sensitize cisplatin-induced apoptosis through the activation of the p53 pathway, therefore represented a synergism effect on aneuploid [MASK] cells.
- We then investigated the effect of [MASK] in combination with [MASK] in cancer cell lines, and we demonstrated a synergistic growth inhibitory effect in GEO and [MASK] cells, evident also with suboptimal doses of [MASK].
- To explore the mechanism of synergistic effects by combining [MASK] and [MASK], we first detected apoptosis by PI staining in [MASK] and H1299 cells that displayed strong synergistic effects in the cytotoxicity assay.
- The combination of 966 and either [MASK], [MASK], or [MASK] led to further reductions in cell growth than either agent alone in [MASK] cells, but these effects were additive, not synergistic.

A.2.2 Iteration 2

- Preclinical [MASK] have indicated a synergistic effect for the [MASK] of anti-EGFR therapy plus [MASK] in breast cancer [MASK].
- As shown in Figure 4G, overexpression of BCL-xL or, to a slightly lesser extent, BCL-xL-Mut1 inhibited the synergistic induction of apoptosis in [MASK] [MASK] [MASK] by the combined [MASK] with [MASK] and [MASK].
- Another study presented that the [MASK] of [MASK] and WZB117 exerts a synergistic cytotoxic effect against breast cancer [MASK].
- In this study, we showed that clinically achievable concentrations of [MASK] and [MASK] in [MASK] could induce a synergistic growth arrest in several human PIK3CA wild-type cancer [MASK].

- The combinative [MASK] of FEN1 inhibitor and 1 nM [MASK] [MASK] a synthetic lethal effect, which synergistically suppressed cancer cell proliferation and significantly mediated apoptosis both in vitro and in vivo.
- [MASK] of LDD1937 and AraC together showed a synergism in the cytotoxic effect on the [MASK] [MASK], and an additive effect was observed between [MASK] and LDD1937.
- [MASK] using [MASK] in [MASK] with other chemotherapeutic drugs or natural compounds with anti-cancer potential may hold the key to identifying synergistic effects of the [MASK] therapy, thereby effectively decreasing the dosage required with better therapeutic efficiency when compared to usage as a monotherapy in the [MASK] of cancer.
- Given that the [MASK] of [MASK] and IMiDs, including [MASK] and pomalidomide, have been shown to result in synergistic apoptotic MM cell death in vitro, the clinical activity of [MASK] could overcome IMiD [MASK] in myeloma patients, and our correlative data justify a phase [MASK] [MASK] of [MASK] and IMiD.
- Interestingly, combinatory [MASK] with 50 muM [MASK] and 25 mM or 40 mM nicotinamide showed a synergistic effect on cell viability, which was observed in both cell [MASK].
- Chou and Talalay analysis of the data suggested that [MASK] of [MASK] and [MASK] was synergistic at a number of selected drug ratios and over a broad range of effective doses.

A.2.3 Iteration 3

- The antioxidant capacity of AXT is 1.5X that of vitamin E. This [MASK] reports on previously unknown findings concerning the synergistic antioxidative effects of combining AXT and HupA, using a previously established [MASK] system to characterize therapeutic agents that can scavenge [MASK] radicals and protect cells from tert-butyl hydroperoxide.
- The resulting [MASK] index theorem of Chou-Talalay was [MASK] to calculate quantitatively whether the [MASK] of drugs [MASK] in an additive, synergistic, or antagonistic effect.
- Recent in vitro [MASK] demonstrate that the [MASK] of H101 and [MASK] exerts a synergistic antitumor effect to uveal melanoma cells without enhanced toxicity to normal cells via a type of cell cycle block, reflecting H101 a promising agent in treating melanoma.
- Interestingly, combinatory [MASK] with 50 muM [MASK] and 25 mM or 40 mM nicotinamide [MASK] a synergistic effect on cell viability, which was observed in both cell lines.
- Given that the [MASK] of [MASK] and IMiDs, including [MASK] and pomalidomide, have been shown to result in synergistic apoptotic MM cell death in vitro, the [MASK] activity of [MASK] could overcome IMiD [MASK] in myeloma patients, and our correlative data justify a [MASK] [MASK] [MASK] of [MASK] and IMiD.
- This technology offers a new [MASK] [MASK] pathway in reason of it 's targeted-specific pharmacodynamics and simplified pharmacokinetics that may improve the therapeutic effect towards tumor [MASK] In [MASK], here we have identified more than thirteen instances of synergism with other drugs that could improve therapeutic regimens, reducing toxicity and side effects.
- Compared with the respective single anticancer action, PDA-NOC-ZnPc12+ nanoparticles [MASK] better anticancer efficacy in tumor-bearing mice, demonstrating the synergistic effect of [MASK] [MASK] with a cell cycle inhibitor and photosensitizer.
- Gong et al., using multiple human PC cells, found that the [MASK] [MASK] with nexrutine and [MASK] [MASK] in significant alterations of proteins in the STAT3/NF-kappaB signaling axis, and growth [MASK] in a synergistic manner.
- The results of the checkerboard [MASK] [MASK] that when colistin was [MASK] in

[MASK] with PFK-158, synergistic activity was observed against the 31 colistin-resistant GNB.

- The Bliss independence method was [MASK] to evaluate synergism between varying [MASK] of [MASK] and [MASK] in sNF96.2 cells, with the highest synergy attained at low doses for each [MASK] in cell proliferation [MASK].

A.3 Dataset Access

The original GDSC-combo dataset can be accessed at https://figshare.com/articles/dataset/Original_screen_drug_combination_data/16843597.

LINCS dataset can be accessed at https://lincsproject.org/LINCS/.

GDSC dataset can be accessed at https://www.cancerrxgene.org/.

CCLE dataset can be accessed at https://sites.broadinstitute.org/ccle/.

NCI-60 dataset can be accessed at https://dtp.cancer.gov/discovery_development/nci-60/cell_list.htm.

All datasets used in this paper are publicly available, and they are used for research purpose only in this paper. Further use of the datasets is subject to their own licenses. The derivated dataset on the GDSC-combo dataset is subject to the original license. All other data can be used for research only.